PB86-140571

Validation Methods for Chemical Exposure and Hazard Assessment Models

(U.S.) Environmental Research Lab., Athens, GA

Dec 85

U.S. Department of Commerce
National Technical Information Service
NTIS

PB86-140571

EPA/600/D-85/297
December 1985

Validation Methods for Chemical Exposure and Hazard Assessment Models

by

Lawrence A. Burns
Environmental Research Laboratory
Athens, GA 30613

ENVIRONMENTAL RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
ATHENS, GA 30613

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA/600/D-85/297		PB86 140577AS

4. TITLE AND SUBTITLE	5. REPORT DATE
VALIDATION METHODS FOR CHEMICAL EXPOSURE AND HAZARD ASSESSMENT MODELS	December 1985
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
Lawrence A. Burns	
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT NO.
Environmental Reseach Laboratory U.S. Environmental Protection Agency Athens, GA 30613	ACULlA
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
Environmental Research Laboratory - Athens, GA Office of Research and Development U.S. Environmental Protection Agency Athens, GA 30613	14. SPONSORING AGENCY CODE EPA/600/01

15. SUPPLEMENTARY NOTES

16. ABSTRACT

Mathematical modesl and computer simulation codes designed to aid in hazard assessment for environmental protection must be verified and validated before they can be used with confidence in a decision-making or priority-setting context. Operational validation, or full-scale testing via an "appeal to Nature" in real-world situations, is usually the most ambiguous and least satisfactory part of the validation process. In most published studies, objective validity criteria are lacking and evaluation of the model is wholly subjective. This need not be the case, however, because acceptable model performance can usually be defined using relatively uncomplicated accuracy and precision criteria. Under these conditions, objective statistical tests can be formulated and executed to provide unambiguous proofs of validity in individual case studies. Such validation cannot, of course, demonstrate the global validity of a model; they merely provide a single instance of a faliure to invalidate. The accumulation of a series of such validations, however, can give model users confidence in the general reliability and veracity of their decision tools.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
RELEASE TO PUBLIC	UNCLASSIFIED	27
	20. SECURITY CLASS *(This page)*	22. PRICE
	UNCLASSIFIED	$ 9.95

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Abstract

Mathematical models and computer simulation codes designed to aid in hazard assessment for environmental protection must be verified and validated before they can be used with confidence in a decision-making or priority-setting context. Operational validation, or full-scale testing via an "appeal to Nature" in real-world situations, is usually the most ambiguous and least satisfactory part of the validation process. In most published studies, objective validity criteria are lacking and evaluation of the models is wholly subjective. This need not be the case, however, because acceptable model performance can usually be defined using relatively uncomplicated accuracy and precision criteria. Under these conditions, objective statistical tests can be formulated and executed to provide unambiguous proofs of validity in individual case studies. Such validations cannot, of course, demonstrate the global validity of a model; they merely provide a single instance of a failure to invalidate. The accumulation of a series of such validations, however, can give model users confidence in the general reliability and veracity of their decision tools.

1. Introduction

Synthetic chemicals contribute significantly to the modern lifestyle and have come to pervade the modern world. Governmental and industrial organizations have begun to respond to public concern for possible damages to human health and natural ecosystems from releases of potentially toxic materials. Large numbers of chemicals, and large numbers of environmental situations, are in need of evaluation.

In consequence, simulation models are today increasingly entrusted with the task of organizing and analyzing great quantities of information, with the objective of producing summary outputs of direct use in regulation and safety evaluations.

The builders and users of these models are rightfully concerned that the models be valid, because decisions based on them have significant economic, public health, and ecological consequences. Substantial intellectual resources have been devoted to constructing a systematic framework for testing operational models (Oren 1981, Mihram 1972, Sargent 1982). Despite the existence of these generalized guidelines, the specifics of model validation remain poorly developed and too frequently entirely subjective (Chapra and Reckhow 1983).

The problem of subjective validity criteria is especially acute in the case of hazard and risk analysis models for toxic chemicals because of a continual conflict in model users', as against model builders', perceptions of the risk and decision structures surrounding validation studies. Validation studies are almost always posed so as to examine the primary hypothesis that "the model is valid," with a consequent focus on minimization of the risk of rejecting a true model (builder's risk), i.e., Type I error. The users of models are concerned only with Type II error, that is, the error and subsequent consequences of accepting an invalid model (user's risk). Because Type II risks are usually not well-controlled, perceptive users of models are subject to continual doubt as to the reliability of their regulatory and evaluative tools. In some instances, validation studies have been designed to control user's risks by creating an especially demanding experimental frame that increases the likelihood of detecting false models (Lassiter, Parrish, and Burns 1985). The efficacy of such designs is always open to post-facto criticism, however, and they do not fully meet the need for wholly objective reporting on the performance and reliability of a simulation model.

This report describes a method for designing objective validation studies that control both Type I and Type II risks, while evaluating the accuracy and precision (bias and dispersion) of simulation models.

Hazard and risk analyses frequently involve the evaluation of chemicals that have not yet been manufactured in significant quantities, and thus have not yet been released into the environment.

Empirical models, which depend upon the existence of relevant data for interpolation to unobserved situations, therefore are not useful in this context. For predictive purposes, mechanistic, process-based causal models are required (Baughman and Burns 1980). For environmental chemicals, these models are constructed to represent the interaction of intrinsic chemical properties (measurable in the laboratory) and environmental factors (measured independently of anthropogenic chemicals) that control the expression of chemical reactivity and toxicity in the biosphere. From a set of valid process models, a larger scale simulation model can be constructed and mechanized on a general-purpose computer. Once the simulation model has been properly verified, its operational validity must be tested by comparing model outputs with observations on the "prototype" system, that is, with the real-world entity of concern (Burns 1983).

2. Goals and Constraints in Validation Studies

The term "null hypothesis" (Ho) is generally, although not of necessity, used to refer to a condition of "no difference." Although statisticians are aware that "validation" is a nonspecific goal that should be more precisely formulated in terms of the purposes of the model and its operational range, they are rarely made aware of the ultimate risk factors associated with the use of simulation models for regulatory purposes. Thus, almost all validation studies are constructed around a null hypothesis that "the model is valid," that is, that the mean values of the model predictions and of the observations on the prototype are not "significantly different" (Balci and Sargent 1981). This approach has several inherent difficulties. First, the observations on the prototype and the measurements of the parameters used to drive the model are both subject to error. When the variance in the parameters propagates into the model predictions, the paradox arises that a less precise model is better able to withstand validation tests (Figure 1). Second, as mentioned above, the

focus of risk control is on the modeler's risk of rejecting a true model (Type I error), and the user's risk is usually ignored. Finally, because no model is an absolutely perfect description of reality, continued testing must ultimately lead to the rejection of this hypothesis. These problems can be eliminated by reformulating the goals of validation studies to better conform with four principles of good scientific practice.

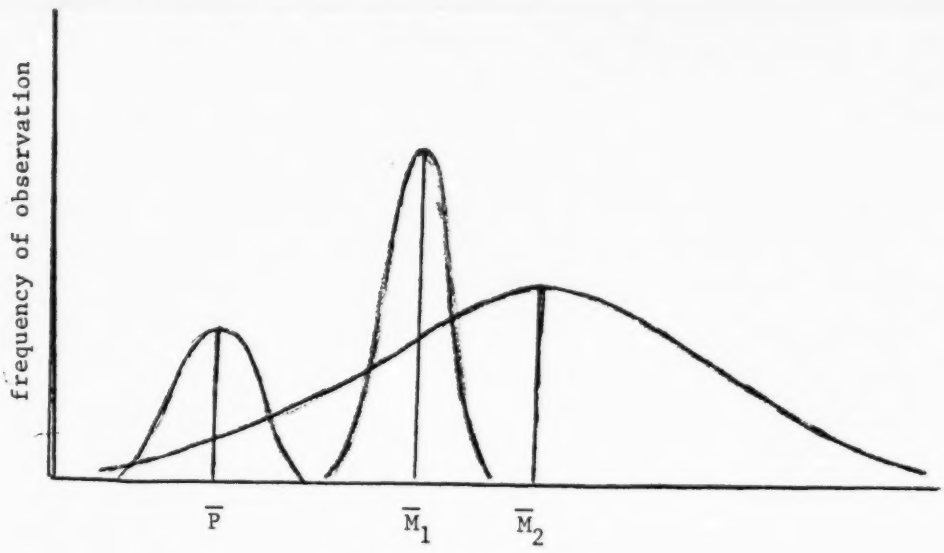


Figure 1. Hypothetical distributions of observations on the prototype P and two models M_1 and M_2. Model M_2 is less accurate and less precise than M_1, but will be perceived as more "valid."

First, when conducting a statistical evaluation of any experimental situation, good practice demands that the null hypothesis be phrased as the opposite of what the experimenter wishes to prove; the tests can then be constructed so as to protect against unconscious bias (stemming from the modeler's natural desire to have constructed a valid model) by making it difficult to reject Ho. In validating simulation models, then, the null hypothesis should always be formulated in terms of Ho: the model is invalid, versus an alternative

hypothesis Ha: the model is valid. (Note that this is the opposite of the usual practice in the field.) This has the advantage of re-assigning the user's risk to Type I error, and the modeler's risk to Type II error. To do this, explicit quantitative performance criteria must be imposed on the model. These criteria are external to the model and to the experimental framework within which a particular validation study is conducted. They serve to define the behavior of the test statistic under the conditions of Ho and to set critical points for its rejection.

Second, whenever possible statistical tests should be formulated so that Ho is a "simple" (as opposed to a "composite") hypothesis, that is, so that the underlying distribution and the values of parameters are completely specified (Freund 1962). Parametric tests should be preferred over nonparametric techniques whenever possible. Nonparametric tests require fewer computations and assumptions about underlying distributions, but they also involve greater probabilities of Type II errors (rejecting a valid simulation model) for a given level of user's risk, a situation highly undesirable from the modeler's perspective.

Third, it must be recognized that both the model and the observations on the prototype involve the same experimental unit, even though both are subject to (often independent) measurement errors. For this reason, only "paired sample" techniques are appropriate to the situation. As was the case for nonparametric analyses, when inappropriate techniqes (e.g., tests of differences between means) are applied, the frequency of Type II errors (modeler's risk) will tend to rise, in this case as a response to increasing variances.

Fourth, given a situation in which parametric tests can be applied, both user's and modeler's risks can be controlled at predetermined levels during the design of the study. Selection of an appropriate sample size can be governed by selection of appropriate risks of Type

I and Type II errors (Diamond 1981), or the level of Type I error can be predetermined, followed by an examination of the risk of Type II error as a function of sample size.

3. A Methodology for Validation of Simulation Models

To apply these principles, it must first be recognized that objective test methods require objective criteria for judgment of model validity. The criteria usually must be based on considerations external to the model, that is, on the needs of the model user for accuracy and precision in model outputs. Chemical exposure models are usually used in a context of toxicological safety evaluations, in which a consistent factor-of-two error is probably not excessive. We, thus, can formulate one objective test of validity as follows.

> "Predicted values from the model must be within a factor of two of reality at least 95 percent of the time."

Paired-sample techniques require that the validation focus on the differences between the model and the prototype on a point-to-point basis. The error in the prediction is a function of the difference between the model prediction "M" and the observation on the prototype "O". These observational pairs are usually taken over a period of time, leading to a series $\{P_1, O_1;\ P_2, O_2;\ \ldots\}$.

The predictions can be arrived at in several ways:

1. When the model parameters are measured at $t_1, t_2 \ldots$, the test is usually called a "structural validation," that is, it tests the mathematical structure of the model.

2. When the model is run with mean values of the parameters (degenerate random variables, i.e., of zero variance), the test may be termed a "parameter validation." It tests the ability of the model to function adequately when its input data are subject to larger errors and approximations.

3. When the model has been designed for Monte-Carlo sampling of the parameter space, parameter validation using degenerate random variables can be followed by Monte-Carlo tests that incorporate stochastic effects into the model predictions. These tests would provide a full-scale parameter validation of models with explicit provision for stochasticity in their (environmental and/or chemical) input parameters.

Regardless of the type of validation test, paired-sample error measures allow for simultaneous stochastic effects on both M (the model predictions) and O (observations on the prototype).

The first test that must be executed is a simple test for correlation between the M and O series. (Notice that regression techniques cannot be rigorously applied because neither M nor O is known precisely.) If M and O are not positively correlated, then the mean value of O would presumably be a better model than is the simulation program. This test can be simply phrased as

$$H_o: \rho = 0$$

vs. $$H_a: \rho > 0$$

where ρ is the correlation coefficient, and testing is conducted at a significance level of 0.01 in order to maximize protection from false models. In designed validation studies, models are unlikely to fail this test because of the wide range of conditions typically selected for testing. In studies of real-world systems, failures are more likely; in any case correlation serves as a convenient screening test and confirmation that continued analysis is warranted.

Implementation of the validity criterion given above requires a measure of the observed error "D" such that a situation of no difference between the M and O gives a value of zero, increasing to a critical value D* at a factor-of-two difference. For parametric

testing, the underlying distribution of the comparison function D should be Normal, and the sampled distribution of D in a specific test should be at least approximately normally distributed. One candidate is the logarithmic transformation, such that

$$D = \log(M) - \log(O)$$

This function has the desired properties:

at M = O, D = 0.0
at M = 2 O, D = +0.301
at M = O/2, D = -0.301

Thus, the maximum permissible bias, for a model with absolute precision (that is, zero variance), would occur at $|D| = 0.301 = D^*$. The validity criterion includes a precision constraint as well: 95% of the errors must be within the limits of permissible bias. This requirement leads to a maximum permissible value for the standard deviation of D, in addition to its maximum mean value.

For a completely unbiased model (the chimerical $\mu_D = 0$), the area under the limits [-0.301,+0.301] may not exceed 0.95 in order that the model be judged valid. For a Standard Normal distribution, 95% of the area under the curve is encompassed by the interval -1.96 to +1.96 in Z, where Z is N(Z;0,1) (that is, Z is normally distributed with mean zero and variance 1). The corresponding points on the underlying distribution of D (assuming it to be Normal) are the limits of permissible bias -0.301 and +0.301. In order to compute the maximum permissible value for σ , the variance of D, we need only observe that when D takes on the value D*, the corresponding standardized random variable Z assumes the value $(X-\mu)/\sigma$. For an unbiased model $(\mu=0)$, the critical value of the standard deviation of D, say σ^* is

$$\sigma^* = D^*/1.96 = 0.154$$

(see Figure 2). This suggests that one test that must be performed during the validation is

$$H_o: \sigma = 0.154$$

vs. $$H_a: \sigma < 0.154$$

using, say, a χ^2 test with alpha = 0.01. Note that the test is phrased with a simple null hypothesis, a one-sided composite alternative, and a test designed to preserve our interest in valid models by making it difficult to reject Ho. Clearly a goodness-of-fit test of the supposition that D is normally distributed will also be required, which can be posed in the conventional way as

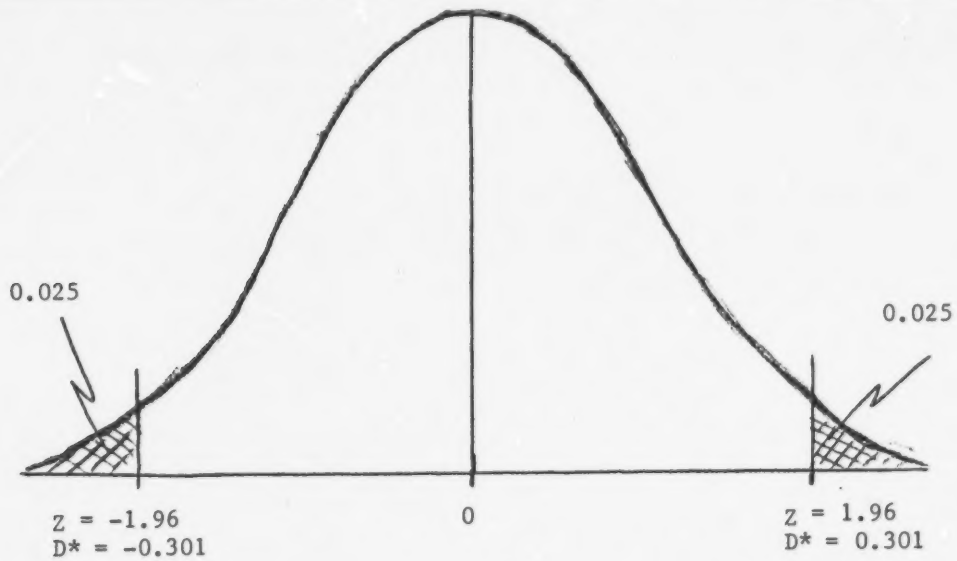
Ho: D is (approximately) Normal

vs. Ha: D is distributed in some other way.

This test also can be conducted via a χ^2 test (or a Kolmolgorov-Smirnoff test) but in this case protection from false models requires that it be somewhat easier to reject Ho. For this example I have chosen to relax alpha to 0.05 (for convenience), in preference to alpha=0.01, in order to continue to minimize model user's risks, in the sense that it is thus easier to reject the premises upon which the analysis is constructed. A minimum of four cells (groups) are required to conduct this test, each of which should conventionally have an expected frequency of five or more. This suggests that sample sizes should not be less than about 20 in order to validate normality assumptions.

The method of constructing sample sizes for controlled levels of user's and modeler's risk (Diamond 1981) requires the specification of acceptable levels of each kind of risk. In this instance, we will select both risk levels at 0.01. Clearly the appropriate values for these risks depend on the perceived social significance of the acceptance and use of false models, and of the rejection of true models;

the methodology does not depend on the particular values selected. This methodology does possess the significant advantage, however, of making the risks explicitly known.



$$\int_{-1.96}^{1.96} N(Z;\ 0,\ 1) = 0.95$$

if D is $N(D;\ \mu, \sigma)$

$$Z = \frac{D-\mu}{\sigma}$$

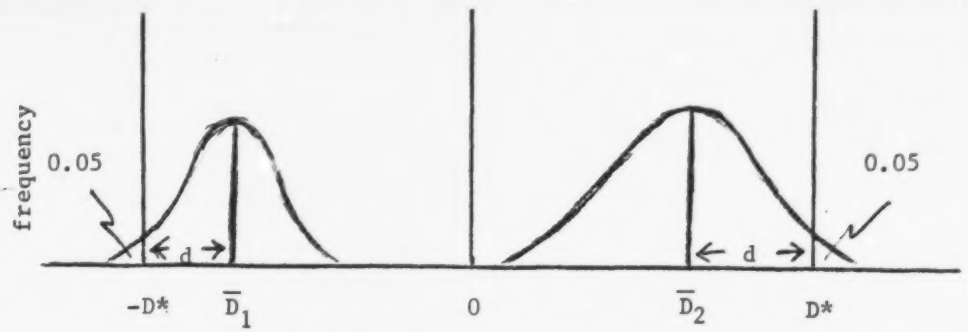
$$\int_{-1.96}^{1.96} N(Z;\ 0,\ 1) = 0.95 = \int_{-D^*}^{+D^*} N(D;\ 0,\ \sigma^*)$$

Figure 2. Limiting values of model bias and dispersion required to meet validity criterion.

The method also requires specification of the least effect that must be detectable in the analysis, which, for paired-sample analyses, must be phrased in terms of the standard deviation (here of the error measure D). This value can be derived by considering the fact that all real models possess some bias, even if it is not detectable (i.e., "statistically significant").

(In fact, the simple presence of detectable bias is not of any particular interest in a validation study, because its presence can be conceded from the outset and because sufficient testing is ultimately certain to detect it. The points of concern are whether the accuracy and precision of the model are sufficient to meet the validity criterion (95% of predictions within a factor of two of reality), and the method of translating this criterion into an objective test based on the bias and dispersion of the error measure.)

The smallest interval of interest will occur when sufficient bias is present to allocate all violations of the validity criterion to one side or the other of the error distribution. The situation is depicted in Figure 3 for a case in which the model is producing predictions somewhat smaller than the observations of the prototype, that is, D is less than zero (the argument is symmetrical for D>0).



$$\int_{-\infty}^{-D^*} N(D_1;\ \mu,\sigma) = 0.05 = \int_{-\infty}^{-1.645} N(Z;\ 0,\ 1)$$

$$\int_{1.645}^{\infty} N\ (Z;\ 0,\ 1) = 0.05 = \int_{D^*}^{\infty} N\ (D_2;\ \mu,\sigma)$$

Figure 3. Definition sketch for formulation of objective validation tests.

Assuming that $\sigma < \sigma^*$ (which assumption will be validated in a separate test), we need concern ourselves only with the left-hand side of the distribution of D. The true value of D (μ_D) must be greater than -D* by a sufficient amount so that only 5% of the total area under the curve is to the left of -D*.

In the Standard Normal distribution, this value corresponds to $Z = -1.645$, and this is the interval "d" that must be detectable in the validation study. (The final test will be posed in terms of single-sided alternatives, with selection of the appropriate test predicated on whether the estimated mean value of D is less than, or greater than, zero.) As was done above for determining the maximum permissible value of σ ($\sigma^* = 0.154$), the value of d can be computed by taking advantage of the relationship between any normally distributed random variable X and the Standard Normal variate Z:

$$Z = (X-\mu) / \sigma$$

As $Z = 1.645$, and $(X-\mu)$ is the distance $D^*-\mu_D$ = the difference "d" that must be detectable in the study, substitution in this formula leads to

$$d = (D^*-\mu_D) = 1.645\ \sigma$$

The formula for computation of the required sample size (Diamond 1981) is

$$N = (U_a + U_b)^2\ \ \sigma^2/d^2$$

where U_a and U_b are the (single-sided) user's and modeler's risk points on the normal probability curve; here both have a value of 2.326 (Diamond 1981:325). Substituting $d = 1.645\ \sigma$ yields a value of $N = 7.997$. The true value of σ is actually unknown, however, so this

estimate of N is too low. The validity tests will be based on the t distribution rather than the normal distribution, and the equation for N must be revised to read

$$N_t = (t_a + t_b)^2 \cdot \sigma^2/d^2$$

Taking N-1 = 6.997 degrees of freedom, the probability point on the t distribution for an 0.01 level of risk is 3.0 (Diamond 1981:326), and

$$N_t = (3 + 3)^2 / 1.645^2 = 13.3$$

Therefore, the collection of 14 samples will control both user's and modeler's risk at a level of 0.01, for a normally distributed error measure having a standard deviation $\sigma < 0.154$.

The actual single-sided tests on the error measure can now be specified. These tests incorporate both the accuracy and the precision requirements of the validity criterion imposed on the model, with the selection of the appropriate test dependent on the sign of the sample mean of the error measure D (Figure 3). In order to phrase the final test as a simple (noncomposite) null hypothesis, we will specify that the model will be rejected as invalid if it is at (or, by implication, beyond) the boundary of the validity criterion. This is a slightly more restrictive test than the original criterion, but its advantages in posing the null hypothesis for risk control greatly outweigh any loss of latitude in acceptable model performance. (The criterion was not initially specified in this form because of the convoluted language required. In its new form, it now reads: "The model must be closer to reality than a factor of two, at least slightly more than 95% of the time.") The appropriate hypotheses are

for $\bar{D} > 0$	for $\bar{D} < 0$
Ho: $\mu_D = (D^*-d)$	$\mu_D = (-D^*+d)$
Ha: $\mu_D < (D^*-d)$	$\mu_D > (-D^*+d)$

where d, the least significant difference, is computed from the sample standard deviation "s" of D, and the t distribution at a probability point of 0.05 (from the validity criterion of 95% precision) and n-1 (= 13 in this design) degrees of freedom.

$$d = t_{0.05,n-1}\ s\ /\sqrt{n}$$

Finally, the validity test is conducted by comparing the value of the t distribution ($t_c = t_{,n-1}$) to the sample t statistic computed from

$$t = \frac{(\bar{D} - (D^*-d))\sqrt{n}}{s} \qquad \text{or} \qquad t = \frac{(\bar{D} - (-D^*+d))\sqrt{n}}{s}$$

in which Ho is rejected if

$$t < -t_c \qquad \text{or} \qquad t > t_c$$

The steps required for a validation study can now be enumerated:

Step 1: Select appropriate validity criteria and establish acceptable levels of user's and modeler's risk. (Except as otherwise noted, upon completion of Step 2 the level of user's risk selected in Step 1 should be used as the alpha risk for all subsequent analyses.)

Step 2: Determine an appropriate error measure for

Ho: the model is invalid for the acceptable range of accuracy and precision in the current context, versus

Ha: the model is valid, at least in this instance.

Step 3: Specify the minimum sample size requirements needed to control risks at acceptable levels and to provide sufficient data for tests of underlying assumptions.

Step 4: Collect point-for-point sample data on the model parameters and the experimental unit (prototype), producing a set of paired samples of simulation model predictions and observations of equivalent properties of the prototype system.

Step 5: Test the paired data for significant correlation; reject the model if they are not positively correlated.

Step 6: Compute the dataset for the error measure D, and test the hypothesis that D is approximately normal (via a χ^2 test at a level of significance suitably less restrictive than the alpha risk level in use to control user's risks, or via a Kolmolgorov-Smirnoff test or a normal probability plot). If the null hypothesis is rejected, the analysis is invalid and will at the least require reformulation.

Step 7: Test the null hypothesis

Ho: $\sigma = \sigma^*$ versus the one-sided alternative

Ha: $\sigma < \sigma^*$ where σ^* is the critical maximum permissible variance derived from the validity criterion imposed in Step 1.

Step 8: Test the validity of the simulation model by comparison of the computed value of the t statistic to the appropriate point on the t distribution; if Ho is rejected then the model can be concluded to be valid (at least in this instance) with certainty of at least $(1-\alpha)$, where "α" is the level of model user's risk selected in Step 1.

4. Example: Photolysis of DMDE in EXAMS II

The Exposure Analysis Modeling System (EXAMS) contains algorithms for computing radiative transfer in the atmosphere as a function of location and climate and for coupling the absorption spectrum and quantum yield of a synthetic chemical to computed spectral irradiance in order to arrive at an estimate of the rate of photochemical reactions in the aqueous environment. (Burns et al. 1982, 1985). Zepp (Zepp and Cline 1977; personal communication) has examined the clear-sky behavior of DMDE (1,1-*bis*(*p*-methoxy phenyl)-2,2-dichloroethylene) at solar noon, collecting 19 independent samples over the course of the calendar year. The EXAMS program calculates rate constants based on mean daily conditions on the day of the month when the declination results in the monthly mean value of the incoming extra-terrestrial irradiance (Iqbal 1983:62). To form a commensurate dataset, the rate constant K computed by the model must be cosine corrected for having been averaged over the course of a day, and then adjusted for daylength:

$$K\ (\text{noon value}) = (\text{daily value})\ (\pi/2)\ \ (24/\text{daylength})$$

The EXAMS program was loaded with the declination and radius vector of the Earth on the dates at which the observations O were taken. The parameter validity of the model was investigated by loading the code via published values of its environmental input parameters. The latitudinal monthly mean values of stratospheric ozone were taken from Iqbal's compilation (Iqbal 1983:89), interpolated to 34° North, the latitude of the laboratory at which the observations were collected. EXAMS' input atmospheric turbidities (Green and Schippnick 1982) were derived from published values of the Beta parameter in the Angstrom formula for atmospheric turbidity. These data were taken from Raleigh-Durham, North Carolina, the nearest station with a published dataset (Iqbal 1983:120). These values are not wholly appropriate to clear-sky conditions, because they are derived from time-series of airport visibility observations, which in turn are presumably contaminated with data taken under heavy cloud cover,

precipitation, and fog. They thus constitute a fairly rigorous test of the model's ability to withstand reasonable levels of error and approximation in its input parameters. The resulting datasets are depicted in Figure 4, and the computed differences (D = log(M) - log (O)) are listed in Table 1.

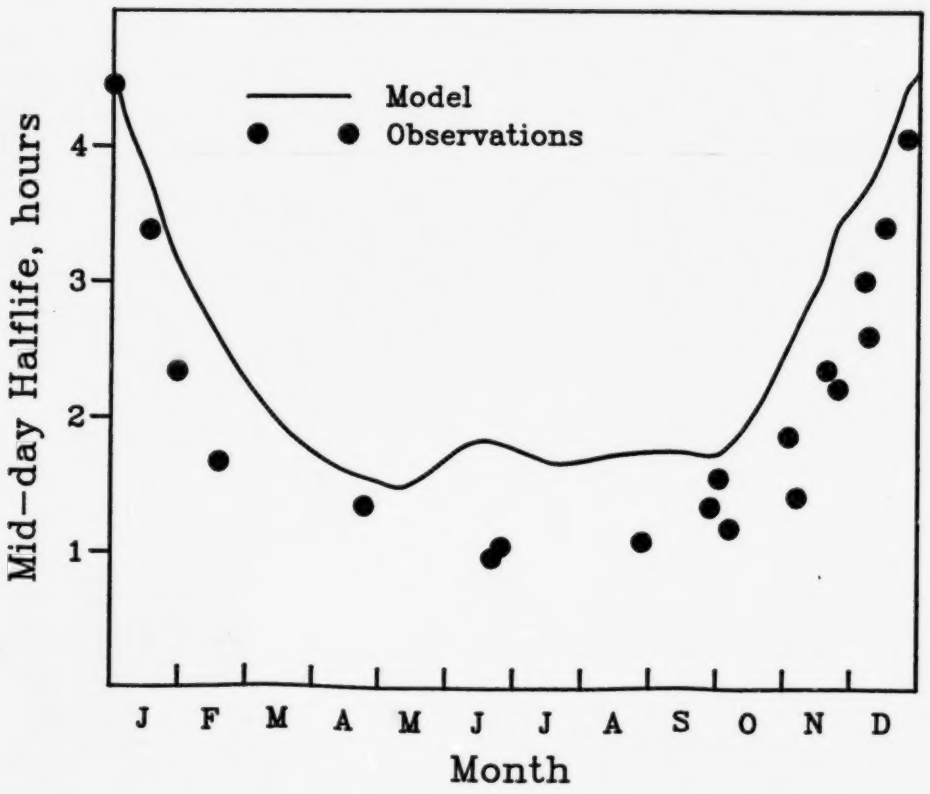


Figure 4. Mid-day photolytic halflife of DMDE under cloudless skies at Athens, Georgia.

Table 1. Observed and Predicted (EXAMS II) Mid-day Halflives of DMDE at Athens, Georgia. Monthly values of Beta (Angstrom turbidity) and latitudinal mean Ozone from Iqbal (1983).

Date	Beta	Ozone cm NTP	Mid-day Halflives (h) Observed	Model	Difference D
2.I	0.065	0.282	4.42	4.37	-0.005
23.I			3.38	3.52	0.018
31.I			2.31	3.12	0.130
19.II	0.083	0.296	1.67	2.76	0.218
24.IV	0.125	0.316	1.36	1.49	0.040
21.VI	0.250	0.312	0.96	1.83	0.280
23.VI			1.02	1.83	0.254
25.VIII	0.230	0.288	1.06	1.81	0.232
26.IX	0.165	0.274	1.32	1.91	0.160
30.IX			1.55	2.01	0.113
2.X	0.100	0.264	1.20	1.69	0.149
31.X			1.87	2.60	0.184
1.XI	0.085	0.268	1.42	2.59	0.261
19.XI			2.35	3.44	0.166
20.XI			2.24	3.46	0.189
3.XII	0.070	0.278	3.00	4.09	0.135
4.XII			2.60	4.11	0.199
12.XII			3.40	4.38	0.110
22.XII			4.05	4.49	0.045
Sample mean ($\bar{X}$)			2.17	2.91	0.152
standard deviation (s)			1.06	1.05	0.083

The first four steps of the validation methodology have been accomplished. (The example computations apply to this study of DMDE photolysis in EXAMS). The four analysis steps (Steps 5-8 above) give the following results.

Step 5. Correlation analysis to validate the assumption that the simulation model is a better predictor of the observations than is the sample mean.

Correlation of the O and P datasets in Table 1 yields a value for the sample correlation coefficient r of 0.92. In testing the null hypothesis (Ho: $\rho = 0$) versus the one-sided alternative (Ha: $\rho > 0$) at a 0.01 level of significance, the computed value of the Z statistic must be greater than $Z_{.01}=2.33$. Computing Z in the usual way (Freund 1962:311) gives Z = 6.44. The null hypothesis can thus be safely rejected, leading to the conclusion that the observations and model predictions are indeed positively correlated, and the validation analysis may proceed.

Step 6. Test the (approximate) normality of the distribution of the error measure D. This step tests the null hypothesis

Ho: D is approximately normal

vs. Ha: D is not normally distributed
and the analysis must be reformulated.

To execute this test, the sample mean and standard deviation are used to predict the frequency distribution of the observations. These expected values are then compared to the actual distribution of the sample errors via a χ^2 "goodness-of-fit" test. The test can be conducted by partitioning the observed data into four cells of equal frequency centered on the sample mean (0.152). The value of Z in the Standard Normal (0.675) giving the 25% points was converted to matching

intervals on the distribution of D, assuming D to be normally distributed with mean 0.152 and standard deviation 0.083.

Interval	Observed Counts	Expected Proportion	Expected Counts
< 0.096	4	0.25	4.75
0.096 - 0.1520	5	0.25	4.75
0.1520 - 0.2082	5	0.25	4.75
> 0.2082	5	0.25	4.75

The χ^2 statistic is computed in the usual way as the sum of the $(O-E)^2/E$, giving in this case a value of 0.43. As this value is less than the (0.05, 1) critical value of χ^2 (3.841), the null hypothesis that the distribution of D is approximately normal can be accepted.

Step 7. This step tests the sample standard deviation for exceedence of the critical value specified in the original validity criterion. The hypotheses are phrased as

$$H_o: \sigma = \sigma^* = 0.154$$

vs. $$H_a: \sigma < 0.154$$

In this test, the null hypothesis can be rejected if the computed χ^2 statistic is less than the value of χ^2 for the 0.01 level of significance with n-1 (18) degrees of freedom ($\chi^2 = 7.015$).

The sample statistic is $\chi^2 = \dfrac{(n-1)\,s^2}{(\sigma)^2} = \dfrac{(18)(0.083)^2}{(0.154)^2} = 5.23$

and, as this value is less than 7.015, it can be concluded with 99% confidence that the precision of the model is adequate.

Step 8: Test the validity of the model, given criteria for acceptable accuracy and precision. Because the sample mean of the error observations is greater than zero, the appropriate null hypothesis is

$H_o: \mu_D = (D^*-d)$ versus $H_a: \mu_D < (D^*-d)$

In this case, $D^* = 0.301$, and

$$d = \frac{t_{.05,18}\, s}{\sqrt{n}} = \frac{(2.101)(0.0833)}{4.36} = 0.0401$$

The critical value t_c is found for 18 degrees of freedom (df) at the 0.01 level of significance ($-t_c = -2.552$). The null hypothesis can be rejected if the computed statistic is less than this value. The computed value is

$$t = \frac{(D - (D^*-d))\sqrt{n}}{s} = \frac{(0.152-(0.301-0.0401))\sqrt{19}}{0.083} = -5.70$$

As $-5.70 < t_c$, the null hypothesis can be rejected, and the simulation model can, with 99% confidence, be accepted as valid.

5. Summary and Conclusions

Given its restricted range of process chemistry, this example validation is an illustration of the fact that global validation can only be accomplished through a series of specific validations. In this instance, however, it has been unambiguously demonstrated that the EXAMS code can compute clear-sky radiative transfer and near-surface photolysis of synthetic chemicals, even under severe

conditions of parameter approximation. Although EXAMS has many additional capabilities, each deserving specific validation, every instance of successful(objective, risk-controlled) validation adds to the confidence that can be placed in this simulation model, via its incremental reduction in global model user's risk.

Objective validations can be conducted only when the criteria for validity are objectively specified. Because the social consequences of accepting false models (inadequate chemical safety regulations) are much more serious than the consequences of rejecting true models (continued research and validation studies), model validations should always be phrased to test the null hypothesis that "the model is invalid." As only a single experimental unit is available for each measurement, paired-sample techniques must be used to create a comparison metric D. When the underlying distribution of D can be identified, the tools of parametric hypothesis testing and statistical decision theory become available to the analyst, as in this report. These tools allow for preliminary experimental designs guaranteed to control user's and modeler's risks at acceptable levels. In addition, the usual tools of statistical inference (confidence limits, etc.) can be used to compare models, and to compute the validity properties exhibited by a simulation model when it is subjected to particular experimental conditions. Even when non-parametric tests must be employed, bowever, the adverse consequences of using false models mandate that model user's risks be assigned to Type I errors by proper formulation of the null hypothesis as "this model is invalid" until proven otherwise.

References

Balci, O., and R.G. Sargent. 1981. A methodology for cost-risk analysis in the statistical validation of simulation models. Communications of the ACM 24:190-197.

Baughman, G.L., and L.A. Burns. 1980. Transport and transformation of chemicals: a perspective. pp. 1-17 In: O.Hutzinger (ed). The Handbook of Environmental Chemistry Vol.2 Part A: Reactions and Processes. Springer-Verlag, Berlin.

Burns, L.A. 1983. Validation of exposure models: the role of conceptual verification, sensitivity analysis, and alternative hypotheses, pp. 255-281 In: W.E. Bishop, R.D. Cardwell, and B.B. Heidolph (eds). Aquatic Toxicology and Hazard Assessment: Sixth Symposium, ASTM STP 802. American Society for Testing and Materials, Philadelphia, Pennsylvania.

Burns, L.A., and D.M. Cline. 1985. Exposure Analysis Modeling System: Reference Manual for EXAMS II. EPA/600/3-85/038, U.S. Environmental Protection Agency, Environmental Research Laboratory, Athens, Georgia.

Burns, L.A., D.M. Cline, and R.R. Lassiter. 1982. Exposure Analysis Modeling System (EXAMS): User Manual and System Documentation. EPA-600/3-82/023, U.S. Environmental Protection Agency, Environmental Research Laboratory, Athens, Georgia.

Chapra, S.C., and K.H. Reckhow. 1983. Engineering Approaches for Lake Management: Mechanistic Modeling. Butterworth Publ., Woburn, Massachusetts.

Diamond, W.J. 1981. Practical Experiment Designs. Lifetime Learning Publications, Belmont, California.

Freund, J.E. 1962. Mathematical Statistics. Prentice-Hall, Englewood Cliffs, New Jersey.

Green, A.E.S., and P.F. Schippnick. 1982. UV-B reaching the surface. pp 5-27 In: J. Calkins (ed). The Role of Solar Ultraviolet radiation in marine ecosystems. Plenum Press, New York.

Iqbal,M. 1983. An Introduction to Solar Radiation. Academic Press, New York.

Lassiter, R.R., R.S. Parrish, and L.A. Burns. 1985. Decomposition by planktonic and attached microorganisms improves chemical fate models. Environmental Toxicology and Chemistry (in press).

Mihram, G.A. 1972. Some practical aspects of the verification and validation of simulation models. Operational Research Quarterly 23:17-29.

Oren, T.I. 1981. Concepts and criteria to assess acceptability of simulation studies: a frame of reference. Communications of the ACM 24:180-189.

Sargent, R.G. 1982. Verification and validation of simulation models. pp. 159-169 In: F.E. Cellier (ed), Progress in Modelling and Simulation. Academic Press, London.

Zepp, R.G., and D.M. Cline. 1977. Rates of direct photolysis in aquatic environment. Environmental Science and Technology 11:359-366.